Technical Progress Report 5/1/94 - 7/31/94
Construction of a Connectionist Network Supercomputer
University of California, Berkeley
ONR URI Grant No. N00014-92-J-1617

1 Abstract

We have made progress in several areas this quarter:

- This quarter saw considerable progress in both high and low level software aspects of the project.
- We have begun work on a CNS-0 system, to be based on the nearly completed T0 processor.
- We have work continued to make progress in the application of analog VLSI to speech pre-processors.

The project continues to have a significant effect on the education of graduate and undergraduate students at our institution. There are currently 16 Ph.D., 1 M.S., and 2 B.S. students associated with the project (some are paid through supporting agencies other than the ONR).

2 Technical Status

2.1 Software and Applications

High-level software. Major progress was achieved in the high-level software section of the project. The Beta release of Sather 1.0 has been successfully ported to a number of platforms and acceptance is very good. The parallel version, pSather, has also made excellent progress; a machine independent run-time interface has been specified and is being implemented. The new implementation also provides support for the ambitious monitoring and debugging system developed by Mark Minas [minas].

An important aspect of the software part of the project is the detailed analysis of the CNS architecture for a variety of problems of interest. In the last quarter we began the study of how the Torrent Architecture and CNS can be applied to image understanding. The results were very encouraging and we will extend these studies under funding from B. Yoon of Arpa. We have also advanced our studies of neural net architectures on CNS. Ben Gomes has completed the first draft of his thesis proposal, which covers the parallel implementation of neural networks.

Low-level software. The work in system software for CNS has concentrated on producing an environment for application and library development. The T0 kernel software and workstation server program have both been refined and now provide a comprehensive environment for running users application programs—this includes full IEEE floating point emulation, a comprehensive range of system calls and hooks for debugging and process monitoring. The debugging hooks have allowed the gdb debugger to be ported to the SPERT system, and additions have been made to support debugging of vector code. To complement the floating point instruction emulation provided by the kernel—which allows arbitrary MIPS architecture floating point code to run—a library of IEEE single precision vector floating point operations is currently undergoing final testing. These will give reasonable high performance (≈15 Mflops) for vectorized code and will be a useful stepping stone when converting applications from scalar floating point to vector fixed point. As in previous quarters, there have been the usual background tasks of fixed point library development and test environment support for the VLSI work.

Speech application. We have started working on speech search algorithms that are more vectorizable than the usual frame-synchronous Viterbi beam search. A number of these are variations of the priority queue based approach that is commonly given the misnomer of "stack decoding". The one we have been working on most recently is an application of simulated annealing to the search. This should be both very vectorizable and very parallelizable. We also have work in progress on N-best approaches, which will also be important for large vocabulary recognition implementations on parallel machines.

2.2 Hardware Development

CNS Systems. This quarter we began detailed design work on a machine we are calling CNS-0. As with the original CNS-1 design, this machine is based on the Torrent processor and is structure in a barrel topology. However, this machine will employ the soon to be completed T0 processor [asanovic]. Unlike the T1 design, this processor does not have the network interface and router integrated on the chip. Rather, the CNS-0 system network will be built using commercially available field programmable logic components. Because this system will be constructed using available components, it will be completed sooner than if we where to wait for another chip design cycle. The CNS-0 construction will allow earlier development of system and application software than would have been possible otherwise. This software will carry over to the CNS-1 system and later CNS systems. Work this quarter has focused in detailed network interface and router design using FPGA components.

In parallel with this effort we have continued the detail design of the CNS-1 machine. Tim Callahan's recently completed Master's thesis reports on the details of the processor–network interface [callahan].

Rambus/communication test chip. Testing of the third version of the communication test chip was successful at 210 MHz. This wraps up an important design area in low voltage

swing signalling methods.

A new test chip, using full voltage swing signalling and a simplified frequency locking technique was designed and is being fabricated at MOSIS now. A new circuit board is also in fabrication, and should allow testing the chips during the next quarter. The full voltage swing signalling method provides several advantages and disadvantages for multiprocessor systems. The primary reason to design and fabricate the new chips is to see whether the disadvantages will be important in a working system.

SPERT board & T0 Chip issues. Detailed simulations of the interface between T0 and the memory on the SPERT board have been completed. Although the memory subsystem was designed early on, several timing parameters could only be estimated due to the incomplete T0 chip design. With T0 nearing tapeout, final adjustments are being made to the clock controller design and the I/O interface on the SPERT board.

2.3 Analog VLSI pre-processors

The analog auditory pre-processor effort has several developments to report this quarter. As previously reported, we have been evaluating our auditory pre-processor chip in a speech recognition task, using a commercial speech recognition toolkit (HTK) as a recognizer; this quarter the evaluation reached its conclusion. For a 200-speaker, telephone-quality, isolated digits database, we found errors rates about 3 to 4 times as high as traditional front-ends (about 97% correct for traditional front ends, about 89% correct for our chip).

In examining the errors made by our chip, we found that most errors occurred in the confusions of starting consonants in a word—for example "five" and "nine" were most often confused. A visual inspection of chip output confirmed the paucity of information for broadband, brief consonants like "f"—this is to be expected in a periodicity-based spectral representation. However, the same chip, loaded with different parameters, can compute other representations that serve as excellent detectors for transient events. Clearly, the use of multiple chips tuned to different representations would lead to better performance—a view shared by many auditory scene analysis researchers.

To facilitate building systems with several auditory pre-processors, we developed an extension to the address-event communications protocol we use in our auditory pre-processors. The extension permits many address-event communications ports to share a common bus, without needing additional chips for bus management. This quarter, we evaluated a test chip for this protocol extension; the scheme worked as expected.

We are now designing a new version of our auditory pre-processor, adding this address-event extension; tapeout is planned for August 31. This chip will allow us to evaluate the use of several different auditory representations simultaneously in our speech recognition application.

3 Presentations

Nelson Morgan, "A Supercomputer for Neural Computation," IEEE International Conference on Neural Networks, June 28-July 2, 1994.

4 Publications

[asanovic] Asanović, K., "T0 Engingeering Data, version 0.11," (This replaces the "T0 Reference Manual"), Internal UCB/ICSI project report.

J. Lazzaro, J. Wawrzynek, and A. Kramer. "Systems Technologies for Silicon Auditory Models". In *IEEE Micro*, (June 1994), 14:3, 7–15.

[callahan] T. Callahan, "Network Interface Specification for the T1 Microprocessor," Master's Thesis Report, Report No. UCB/CSD-94-823, May 1994.

Feldman, J.A.: "Universal High Performance Computing - We Have Just Began," April 1994. To appear in ACM book on HPCC.

[minas] Minas, M.: "Fault Detection for Sequentially Controlled Machines Using Temporal Constraint Nets," in *Preprints of the IFAC Symposium on Fault Detection, Supervision and Safety for Technical Processes (Safeprocess '94)*, pp. 323-328, Helsinki, June 13-16, 1994.

Szyperski, C., Omohundro, S., and Murer, S.: "Engineering a Programming Language: The Type and Class System of Sather", in Gutknecht, J. (ed.) *Programming Languages and System Architectures, Lecture Notes in Computer Science Volume 782*, pp. 208-227, Springer-Verlag, Berlin, 1994; accepted for publication in *Conference on Programming Languages and System Architecture*, Zurich, March 1994; ICSI Technical Report TR-93-064.

Asanović, K., Beck, J., Feldman, J., Morgan, N., and Wawrzynek, J., "A Supercomputer for Neural Computation" Proceedings of the ICNN, vol. 1, pp.5-9, 1994,